Tighter guarantees for the compressive multi-layer perceptron

Kaban, Ata; Thummanusarn, Yamonporn

DOI:
10.1007/978-3-030-04070-3_30

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Kaban, A & Thummanusarn, Y 2018, Tighter guarantees for the compressive multi-layer perceptron. in D Fagan, C Martín-Vide, M O'Neill & M A. Vega-Rodríguez (eds), Theory and Practice of Natural Computing: 7th International Conference, TPNC 2018 Dublin, Ireland, December 12–14, 2018 Proceedings. Lecture Notes in Computer Science, Springer, pp. 388-400, 7th International Conference on the Theory and Practice of Natural Computing (TPNC 2018), Dublin, Ireland, 12/12/18. https://doi.org/10.1007/978-3-030-04070-3_30

Tighter Guarantees for the Compressive Multi-layer Perceptron

Ata Kabán and Yamonporn Thummanusarn

School of Computer Science, University of Birmingham
Edgbaston, B15 2TT, Birmingham, UK
a.kaban@cs.bham.ac.uk, yxt653@cs.bham.ac.uk

Abstract. We are interested in theoretical guarantees for classic 2-layer feed-forward neural networks with sigmoidal activation functions, having inputs linearly compressed by random projection. Due to the speedy increase of the dimensionality of modern data sets, and the development of novel data acquisition devices in compressed sensing, a proper understanding of are the guarantees obtainable is of much practical importance. We start by analysing previous work that attempted to derive a lower bound on the target dimension to ensure low distortion of the outputs under random projection, and we find a disagreement with empirically observed behaviour. We then give a new lower bound on the target dimension that, in contrast with previous work, does not depend on the number of hidden neurons, but only depends on the Frobenius norm of the first layer weights, and in addition it holds for a much larger class of random projections. Numerical experiments agree with our finding. Furthermore, we are able to bound the generalisation error of the compressive network in terms of the error and the expected distortion of the optimal network in the original uncompressed class. These results mean that one can provably learn networks with arbitrarily large number of hidden units from randomly compressed data, as long as there is sufficient regularity in the original learning problem, which our analysis rigorously quantifies.

Keywords: Error analysis · Random projection · Multi-layer perceptron

1 Introduction

Let $\mathcal{X} \subset \mathbb{R}^d$ be an input domain. We denote by $\mathcal{H} = \{x \to h(x) : x \in \mathcal{X}\}$ the function class that implements neural networks of the following parametric form:

$$h(x) = u + \sum_{i=1}^{M} v_i \sigma(w_i^T x) \tag{1}$$

where $\sigma : \mathbb{R}^d \to [-b, b]$ is a Lipschitz continuous bounded activation function – traditionally a sigmoidal function, such as $\sigma(u) = \tanh(u)$, or the logistic function $\sigma(u) = \frac{1}{1+e^{-u}}$. Further, $w_i \in \mathbb{R}^d, u, v_i \in \mathbb{R}$, are weights or parameters of the

network. In practice, these parameters are estimated from a finite set of labelled training points denoted by $\mathcal{T}^N = \{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \overset{\text{i.i.d}}{\sim} \mathcal{D}$, and $\mathcal{D}$ is an unknown distribution over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = [-b, b] \subset \mathbb{R}$.

Now, suppose that d is too large to work with directly – as this is indeed the case in many modern data sets – and we employ random projection (RP) to reduce dimension before feeding the data to the neural network. One of the practical motivations for this approach is the prospect of making use of novel data acquisition techniques from compressed sensing, such as CCD and CMOS cameras [14]. These devices bypass the need to store and process large data sets and instead collect a random linear projection of the data directly. As a result, there has been a surge of interest in studying compressive learning – see e.g. [13] for a recent account.

Denote the random projection (RP) matrix by $R \in \mathbb{R}^{k \times d}$, $k < d$, with independent and identically distributed (i.i.d.) entries drawn independently of $\mathcal{T}^N$, from a suitable 0-mean $1/k$-variance distribution, and the compressed training set is $\mathcal{T}_R^N = \{(Rx_n, y_n)\}_{n=1}^N$. The distribution of the entries of R is usually chosen so as to satisfy the Johnson-Lindenstrauss property [7], e.g. Gaussian or subgaussian. Our results will hold for more general random matrices, as the proof technique we will employ only requires i.i.d. entries from a symmetric 0-mean distribution with finite first four moments.

2 Previous work

The work of [14] studied the problem of dimensionality reduction by Gaussian random projection in two-layer feed-forward networks of the form defined in eq. (1). More precisely, the authors bounded the absolute difference of the outputs of the network before and after random projection, and derived a lower bound on the required target dimension k to ensure low distortion of the outputs on the sample. The statement of the main result of [14] is the following.

Theorem 1 ([14]). *Consider feed-forward neural networks with sigmoidal activation functions and M hidden units. Define $C = L_\sigma \max_i |v_i| \|w_i\| \max_{x \in \mathcal{X}} \|x\|$, where $L_\sigma > 0$ is Lipschitz constant depending on an activation function σ. For any $\eta \in (0, 1)$ and $\delta > 0$, if the dimension after projection k is selected as:*

$$k \geqslant \frac{12C^3(\log M + \log N + \log 2 - \log \delta)}{3C\eta^2 - 2\eta^3}, \tag{2}$$

then, $Pr\{\frac{1}{N}\sum_{n=1}^N |h(x_n) - h(R^T R x_n)| > \eta\} < \delta$, where, $x_i \in \mathcal{X}, i = 1, \ldots, N$, is a given set of N points.

In eq. (2) we observe a logarithmic dependence of the target dimension k on M and N. These seem like mild dependencies, but one wonders whether they are necessary.

However, looking closer, unfortunately we observe a typo in the third line of proof of their Theorem 1 [14], which carries forward and makes its way into the

main statement – that is, the above Theorem 1 is actually incorrect. The issue is the following.

Let $|Z_i|$ denote $|v_i||\sigma(w_i^T R^T Rx) - \sigma(w_i^T x)|$. Lines 2 to 3 of the proof of Theorem 1 (that is Theorem 1 in [14]) are equivalent to the following:

$$\forall \eta > 0, Pr\{\sum_{i=1}^{M} |Z_i| > \eta\} \leqslant \sum_{i=1}^{M} Pr\{|Z_i| > \eta\} \tag{3}$$

To see why eq. (3) is incorrect, we construct a counterexample. Let $M = 2$ and consider the event of rolling two fair dice whose faces hold values from 0.1 to 0.6 for the sake of the argument, and set $\eta = 0.6$. So the left hand side (LHS) of equation 3 represents the probability that the sum of outcomes from the two dice is strictly greater than 0.6, which equals $\frac{21}{36} = \frac{7}{12}$. The right hand side (RHS) represents the sum of the probabilities that the outcome of one dice is strictly greater than 0.6, which equals to 0. This is a contradiction, since $\frac{7}{12} \nleq 0$.

The mistake occurred by missing the denominator M under η when using the union bound inequality. Next, we will correct this, and re-derive the lower bound on k, to obtain a corrected version of eq. (2) in Theorem 1.

2.1 A correction to the previous work

Applying the union bound inequality correctly, for any $\eta > 0$ and any x we have $Pr\left\{\left|\sum_{i=1}^{M} v_i \left[\sigma(w_i^T R^T Rx) - \sigma(w_i^T x)\right]\right| > \eta\right\} \leqslant \ldots$

$$\leqslant Pr\left\{\sum_{i=1}^{M} |v_i| \left|\sigma(w_i^T R^T Rx) - \sigma(w_i^T x)\right| > \eta\right\} \tag{4}$$

$$\leqslant \sum_{i=1}^{M} Pr\left\{|v_i| \left|\sigma(w_i^T R^T Rx) - \sigma(w_i^T x)\right| > \frac{\eta}{M}\right\} \tag{5}$$

$$\leqslant M \max_i Pr\left\{\left|\sigma(w_i^T R^T Rx) - \sigma(w_i^T x_n)\right| > \frac{\eta}{|v_i| M}\right\} \tag{6}$$

Note M in the denominator, which was missed in the original proof of [14].

Carrying on from this, by the assumption that σ is Lipschitz continuous we have $|\sigma(t + a) - \sigma(a)| \leqslant L_\sigma |a|, t, a \in \mathbb{R}$, where $L_\sigma > 0$ is the Lipschitz constant of the activation function σ. Thus, we have:

$$\left|\sigma(w_i^T R^T Rx) - \sigma(w_i^T x)\right| \leqslant L_\sigma \left|w_i^T R^T Rx - w_i^T x\right| \tag{7}$$

Then, we can bound the probability of the unlikely event that the difference between the dot product of the vectors w_i and x before and after random projection is larger than $\epsilon \|w_i\| \|x\|$, by using for instance Lemma 1 of [14] – this lemma works for Gaussian random projection only, because their proof heavily relies on the rotation-invariance of Gaussians – or Theorem 2.1. from [9] – which applies to the larger class of sub-Gaussian random projections, and has essentially

the same form. The advantage of sub-Gaussian RPs is a better computational scaling while they enjoy similar guarantees [1, 11].

Either way, the Johnson-Lindenstrauss type bound for dot products that we need to apply to eq. (7) is the following:

$$Pr\left\{\left|\left(\frac{w_i^T R^T Rx}{\|w_i\|\|x\|}\right)-\left(\frac{w_i^T x}{\|w_i\|\|x\|}\right)\right|>\epsilon\right\} = Pr\left\{\left|w_i^T R^T Rx - w_i^T x\right| > \epsilon\|w_i\|\|x\|\right\}$$
$$\leqslant 2\exp\left[-k\left(\frac{\epsilon^2}{4}-\frac{\epsilon^3}{6}\right)\right]$$

where $\epsilon \in (0,1)$.

Let $C = L_\sigma \max_i |v_i|\|w_i\| \max_{n\in\{1,\ldots,N\}} \|x_n\|$ and let $\epsilon = \frac{\eta}{C\cdot M}$. Hence, we obtain the bound:

$$Pr\left\{\left|\sum_{i=1}^{M} v_i\left[\sigma(w_i^T R^T Rx)-\sigma(w_i^T x)\right]\right|>\eta\right\} \leqslant 2M\exp\left[-k\left(\frac{\eta^2}{4C^2M^2}-\frac{\eta^3}{6C^3M^3}\right)\right] \quad (8)$$

Since we have N points in total, using the union bound inequality over these, as in [14], brings a factor of N to the right hand side of eq. (8). Finally, choosing the risk tolerance $\delta > 0$, this yields the following lower bound on the required dimension, k:

$$2NM\exp\left[-k\left(\frac{\eta^2}{4C^2M^2}-\frac{\eta^3}{6C^3M^3}\right)\right] \leqslant \delta,$$

Solving for k we obtain:

$$k \geqslant \frac{12C^3M^3(\log M + \log N + \log 2 - \log\delta)}{3CM\eta^2 - 2\eta^3} \quad (9)$$

Eq. (9) is the correct lower bound on k from the correct application of the proof technique of [14]. That is, eq. (9) replaces eq. (2) in the above Theorem 1. We observe however that, after this correction, the dependence of k on M became even stronger. In fact, it is not difficult to show that the right hand side of eq. (9) is of order $\Omega(M^2(\log(MN)))$.

2.2 Numerical checks

Before going further, empirical checks can be useful to gain insights. We are interested to see whether the strong dependence on M that appears in the bound is actually observed empirically.

In the first experiment, we fix $k = 20$. We generate a $d = 100$ dimensional input vector x randomly and then fixed it. We vary the network size M, and for each value tested, we do 100 independent repetitions of the following: Generate parameter values W, v, u randomly (from standard normal) and normalise them to have $\|W\|_{Fro} = 1, \|v\|_2 = 1$. Here, $\|\cdot\|_{Fro}$ denotes the Frobenius norm and $\|\cdot\|_2$ is the l_2 norm. Fixing these, we then generate a random projection (RP)

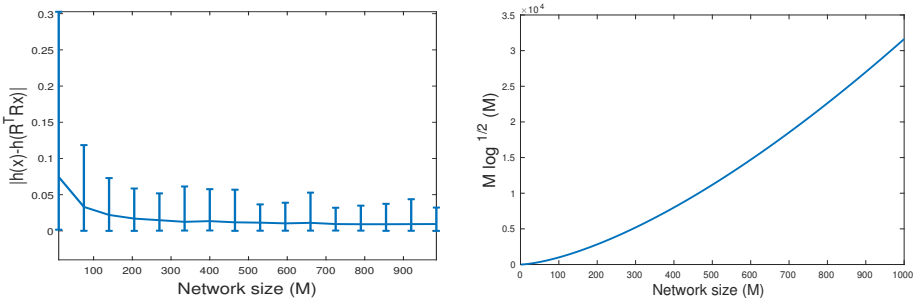

Fig. 1. *Left*: Numerical experiments from 100 independent repetitions, in $d = 100$ dimensions as the network size (M) grows, while keeping $\|W\|_{Fro} = 1$ and $\|v\|_2 = 1$ constant. The target dimension was fixed to $k = 20$. The error bars span the minimum to the maximum of the observed distortions, the continuous line is the average distortion. *Right*: The graph of $M\sqrt{\log M}$, i.e. the order of distortion for a fixed k from the existing analytic bound. From these two plots the disagreement between the bound and the empirical behaviour is most apparent.

matrix and compute the distortion $|h(x) - h(R^T Rx)|$. The left plot in Figure 1 shows the observed distortion values (average, minimum, maximum). From Figure 1 we observe no growth as M increases.

On the other hand, from eq. (9) it follows that the distortion bound η is of order $\Omega(M\sqrt{\log(MN)})$. For comparison, and as a test of the explanatory value of the existing bound, we show on the right hand plot the graph of the function $M\sqrt{\log(M)}$. From these two plots we observe a clear disagreement between the empirical behaviour of the required k and the bound as M varies. Next, with

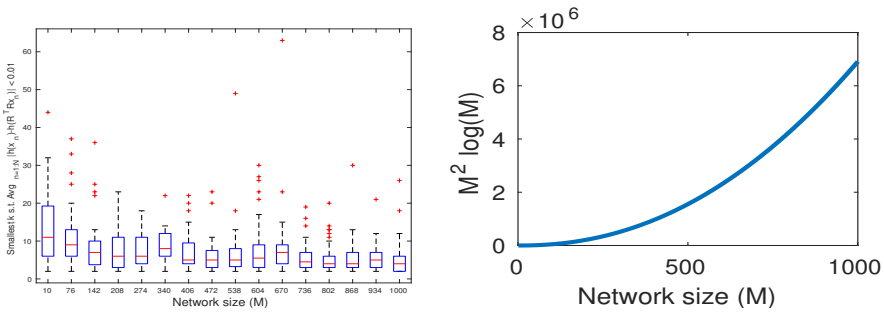

Fig. 2. *Left*: The empirical distribution of the target dimension (k) required for the observed distortion to be below 0.01. *Right*: The graph of $M^2 \log M$. Again, the disagreement between the bound and the empirical behaviour is most apparent.

the same protocol, for each network size and each of the 100 repetitions we

generate RP matrices with different target dimensions k and select the smallest k for which $|h(x) - h(R^T Rx)| \leq \eta \leq 0.01$. If no k satisfied this threshold then we discarded the experiment. Of the successful ones, we plot in Figure 2 the empirical distributions of k on the left hand plot. Side by side, we also plot the function $M^2 \log(M)$, that is the order of the lower bound on k from eq. (9). Again, we observe a disagreement between the empirical behaviour and the analytic bound as M varies, since the bound grows with M while the empirical behaviour appears to be unaffected. Of course, numerics cannot substitute for a proof, but they strongly suggest that it is worth trying a different approach. For that we need different proof ideas. This is the subject of the next section.

3 A better approach

In this section we attack the problem differently. Denote $D := \frac{1}{N}\sum_{n=1}^{N} |h(R^T Rx_n) - h(x_n)|$ the distortion under RP. We have:

$$
\begin{aligned}
D &\leq \frac{1}{N}\sum_{n=1}^{N} |\sum_{i=1}^{M} v_i \left(\sigma(w_i^T R^T Rx_n) - \sigma(w_i^T x_n)\right)| \\
&\leq \|v\|_2 \frac{1}{N}\sum_{n=1}^{N} \sqrt{\sum_{i=1}^{M} [\sigma(w_i^T R^T Rx_n) - \sigma(w_i^T x_n)]^2}, \text{ by Cauchy-Schwartz} \\
&\leq \|v\|_2 \frac{1}{N}\sum_{n=1}^{N} \sqrt{\sum_{i=1}^{M} L_\sigma^2 \cdot [w_i^T R^T Rx_n - w_i^T x_n]^2}, \text{ by Lipschitzness of } \sigma(\cdot) \\
&= L_\sigma \cdot \|v\|_2 \cdot \frac{1}{N}\sum_{n=1}^{N} \|W^T R^T Rx_n - W^T x_n\|_2 \qquad (10)
\end{aligned}
$$

where the matrix $W \in \mathbb{R}^{d \times M}$ has the weight vectors w_i in its columns, and $v = (v_1, ..., v_M)$.

We are able to bound the expectation of this w.r.t. R. After that we will use tail bounds to obtain a high probability bound on D.

To bound the expectation, we need the following lemma, which can be proved using Lemma 2 from [8], and holds for a very general class of R, a class that is larger than sub-Gaussians.

Lemma 2. *Let R be $k \times d$ random matrix with i.i.d. entries drawn from a 0-mean symmetric distribution with finite first four moments, and denote $\kappa_+ = \max(0, \kappa)$ where $\kappa = \frac{E[R_{ij}^4]}{E[R_{ij}^2]^2} - 3$ is the excess kurtosis of the entries.. Then,*

$$
E_R \|W^T R^T Rx - W^T x\| \leq \sqrt{\frac{2 + \kappa_+}{k}} \cdot \|W\|_{Fro} \cdot \|x\|_2. \qquad (11)
$$

For Gaussian the excess kurtosis is 0. So for all sub-Gaussians, $\kappa_+ = 0$.

Applying this Lemma 2 to eq. (10), we obtain:

$$\mathrm{E}_R[D] \leq L_\ell \cdot L_\sigma \cdot \|v\|_2 \cdot \sqrt{\frac{2+\kappa_+}{k}} \cdot \|W\|_{Fro} \cdot \frac{1}{N}\sum_{n=1}^{N} \|x_n\|_2 \tag{12}$$

Now it remains to plug this into a tail bound. By Höffding inequality we have for any $\epsilon > 0$, $\Pr\{D \geqslant \mathrm{E}_R[D] + \epsilon\} \leqslant \exp\left(-2\epsilon^2/\bar{\ell}^2\right)$. Hence, w.p. at least $1-\delta$

$$D \leqslant \mathrm{E}_R[D] + \bar{\ell}\sqrt{\frac{1}{2}\log\frac{1}{\delta}} \tag{13}$$

Note that whenever $\mathrm{E}_R[D]$ is small (close to 0) then so is $D - \mathrm{E}_R[D]$, since D is always positive. A Markov inequality will capture this: $\Pr\{D \geq \epsilon\} \leq \frac{\mathrm{E}_R[D]}{\epsilon}$, and yields w.p. $1-\delta$

$$D \leq \mathrm{E}_R[D] + \frac{1-\delta}{\delta}\mathrm{E}_R[D] \tag{14}$$

Eq.(13) is tighter at small δ, while eq.(14) tightens with decreasing $\mathrm{E}_R[D]$. Taking the minimum we have:

$$D \leq \mathrm{E}_R[D] + \min\left\{\frac{1-\delta}{\delta}\mathrm{E}_R[D], \bar{\ell}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\} \tag{15}$$

Combining eq. (15) with eq. (12) completes the high probability bound on the distortion D.

Most interestingly, and importantly, observe that in this bound there is no direct dependence on either M or N.

3.1 New lower bound on k

We are now in a position to deduce a new lower bound on the required dimension k that keeps the distortion small. Observe that $\mathrm{E}_R[D]$ is a decreasing function of k, so we require that this term is below some threshold $\eta > 0$, yielding the following.

Theorem 3. *For the class of RP matrices R from Lemma 2, the required target dimension of the compressed space that ensures $E_R[D] \leq \eta$ is lower bounded as the following:*

$$k \geq \eta^{-2} \cdot L_\ell^2 \cdot L_\sigma^2 \cdot \|v\|_2^2 \cdot \|W\|_{Fro}^2 \cdot (2+\kappa_+) \cdot \left(\frac{1}{N}\sum_{n=1}^{N}\|x_n\|_2\right)^2 \tag{16}$$

Observe again, there is no dependence on either M or N.

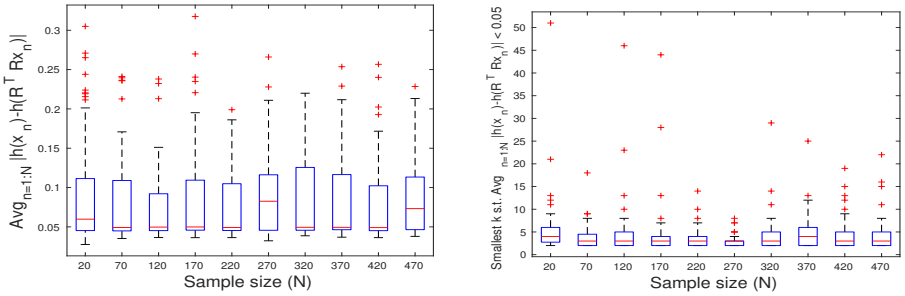

Fig. 3. Numerical experiments from 100 independent repetitions in $d = 100$ dimensions, as the sample size (N) grows, while keeping $\|W\|_{Fro} = 1$ and $\|v\|_2 = 1$ constant. The network size was set to $M = 200$. For the left hand plot, $k = 5$. We observe no increase of the quantities of interest with N.

3.2 Further numerical checks

Figures 1 and 2 from Section 2.2 corroborate our theoretical findings, as indeed we see no increase in distortion or in the required k as the network size M grows – in agreement with our theory. Before concluding, we also check for dependencies on N. For these experiments we fix the network size to $M = 200$, and follow the same protocol as before. For each sample size tested, we do 100 independent repetitions, computing the distortion, and the left hand plot of Figure 3 shows the distribution of these. In the right hand plot we show the distribution of the smallest k that achieves a distortion of less that 0.05. We observe no dependence of the distortion or of the required k as N varies. Based on this, together with the empirical results from Section 2.2 we conclude that our theory agrees with empirical behaviour.

4 Generalisation error of the compressive multi-layer perceptron

So far we have only looked at the distortion of the outputs of the network *on the sample*. In this section we give a bound on the *generalisation error* of the compressive network.

We denote by $\mathcal{H} = \{x \to h(x) : x \in \mathcal{X}, \|v\|_1 \leq C_v\}$ the function class that implements neural networks of the parametric form given in eq. (1), where $C_v > 0$ is a constant.

The training set is denoted as $\mathcal{T}^N = \{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \overset{\text{i.i.d}}{\sim} \mathcal{D}$, where $\mathcal{D}$ is an unknown distribution, $\mathcal{Y} = [-b, b] \subset \mathbb{R}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ be a bounded loss function, assumed to be L_ℓ-Lipschitz in its first argument. The loss function measures the mismatch between the true and the predicted labels of a labelled point. Typically the loss function depends on its arguments only through their product $y \cdot h(x)$ (for classification), or through their difference, $y - h(x)$ (for regression).

It is well known that this class of neural networks (NNs) is capable of approximating any smooth target function arbitrarily well when provided with a sufficient number of hidden units [5] – that is, when the size of the network, M, is large enough. It is also known that, for good generalisation, the size of the weights matters more than the size of the network [3]. So our findings from the previous section seem very relevant.

The quantity of ultimate interest that quantifies the success of learning is the generalisation error (or risk). For an $h \in \mathcal{H}$, this is defined as $\mathrm{E}[\ell \circ h] := \mathrm{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$. Because $\mathcal{D}$ is unknown, we only have access to the empirical error, defined as $\hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ h] = \frac{1}{N}\sum_{n=1}^{N} \ell(h(x_n), y_n)$. The optimal learner within $\mathcal{H}$ will be denoted as $h^* = \arg\inf_{h\in\mathcal{H}} \mathrm{E}[\ell \circ h]$. The sample error minimiser of the loss is $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ h]$.

In the reduced k-dimensional space we have analogous definitions, and will use a subscript to refer to this reduced space. The functions in the reduced space have the form:

$$h_R(Rx) = u_R + \sum_{i=1}^{M} (v_R)_i \cdot \sigma((w_R)_i^T Rx) \tag{17}$$

where $(w_R)_i \in \mathbb{R}^k, u_R, (v_R)_i \in \mathbb{R}$ are the parameters that are estimated from $\mathcal{T}_R^N$. Thus, the compressed function class of our interest is $\mathcal{H}_R = \{Rx \to h_R(Rx) : x \in \mathcal{X}, \|v_R\|_1 \leq C_v\}$ where $C_v > 0$ is a constant. We will not restrict the norms of the nonlinear layer's parameter vectors $(w_R)_i$ because the complexity on this layer is already reduced by the RP.

We will refer to the sample error minimiser in this reduced class as $\hat{h}_R = \arg\min_{h_R\in\mathcal{H}_R} \hat{\mathrm{E}}_{\mathcal{T}_R^N}[\ell \circ h_R]$, where $\hat{h}_R \in \mathcal{H}_R$ and $\hat{\mathrm{E}}_{\mathcal{T}_R^N}[\ell \circ h_R] = \frac{1}{N}\sum_{n=1}^{N} \ell(h_R(Rx_n), y_n)$ is the empirical error in the reduced space. Likewise, the optimal learner within $\mathcal{H}_R$ is denoted as $h_R^* = \arg\min_{h_R\in\mathcal{H}_R} \mathrm{E}[\ell \circ h_R]$.

With these preliminaries in place, the next subsection bounds the generalisation error of 2-layer networks trained by empirical risk minimisation (ERM) on randomly projected data.

4.1 Generalisation error bound

For convenience we assume that $\mathcal{H}$ is closed under negation, hence so is $\mathcal{H}_R$ – that is, $\mathcal{H}_R = -\mathcal{H}_R$. We shall make use of Rademacher complexities [4, 12]; for $\mathcal{H}_R$, the function class of our interest, recall this is defined as the following:

$$\hat{\mathcal{R}}_N(\mathcal{H}_R) = \mathrm{E}_\gamma[\sup_{h_R\in\mathcal{H}_R} \frac{1}{N}\sum_{n=1}^{N} \gamma_n h_R(Rx_n)] \tag{18}$$

where $\gamma = (\gamma_1, ..., \gamma_N)$ and γ_n takes values in $\{-1, 1\}$ with equal probability.

Theorem 4. *Let $\mathcal{H}$ be the function class of feed-forward neural networks of the form defined in eq.(1), with L_σ-Lipschitz continuous activation functions $\sigma : \mathbb{R} \to$*

$[-b, b]$, and assume $\mathcal{H} = -\mathcal{H}$. Let $h^* = \arg\inf_{h\in\mathcal{H}} E[\ell \circ h]$ be the optimal network in this class, with parameters $(W^* = [w_1^*, ..., w_M^*] \in \mathbb{R}^{d\times M}, v^* = (v_1^*, \ldots, v_M^*) \in \mathbb{R}^M$, $u^* \in \mathbb{R})$, where $\|v^*\|_1 \leq C_v$, and $C_v > 0$ is a constant.
Let $\mathcal{T}_R^N$ denote the RP-ed training set of size N, where the original sample is $\mathcal{T}^N \overset{i.i.d}{\sim} \mathcal{D}$, and we assume $E_{x\sim\mathcal{D}}[\|x\|_2] < \infty$. The RP matrix $R \in \mathbb{R}^{k\times d}$, $k \leq d$ has entries R_{ij} drawn i.i.d. from a symmetric distribution with 0-mean and finite first four moments, and let $\kappa_+ = \max\left\{0, \frac{E[R_{ij}^4]}{E[R_{ij}^2]^2} - 3\right\}$. Denote by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ a loss function assumed to be L_ℓ-Lipschitz in its first argument, and let $\hat{h}_R$ be the sample error minimiser of this loss with respect to $\mathcal{T}_R^N$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$, the generalisation error of $\mathcal{H}_R$ is upper bounded as the following:

$$E_{x,y}[\ell \circ \hat{h}_R] \leq E_{x,y}[\ell \circ h^*] + cL_\ell b(1 + C_v) \cdot \sqrt{\frac{k}{N}} + 4\bar{\ell}\sqrt{\frac{\log\frac{2}{\delta}}{2N}}$$
$$+ \left(g_k(W^*, v^*) + \min\left\{ \frac{1-\delta}{\delta} g_k(W^*, v^*), \bar{\ell}\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)} \right\} \right) \mathbf{1}(k < d)$$

where c is an absolute constant, $\mathbf{1}(\cdot)$ is 1 if its argument is true and 0 otherwise, and

$$g_k(W^*, v^*) \leq L_\ell L_\sigma \|v^*\|_2 \cdot \|W^*\|_{Fro} \cdot \sqrt{\frac{2 + \kappa_+}{k}} \cdot E[\|x\|_2].$$

The proof is rather lengthy and is therefore deferred to the full version. At a high level, the generalisation error in the compressed space is decomposed into a distortion term and the complexity of the function class over k-dimensional inputs. As we intentionally did not constrain the first layer weights, we estimate the Rademacher complexity through a fat-shattering bound [2] and Dudley inequality [6], exploiting boundedness of the activation functions, rather than the l2 geometry. This has the additional advantage that $\mathcal{X}$ needs not be bounded.

The meaning of the function $g_k(W^*, v^*)$ in Theorem 4 is very similar to the expected distortion that we analysed in the previous sections, with 3 differences: First, rather than some arbitrary parameters W, v here we have the parameters of the best performing network in the function class, W^*, v^*. This is desirable because we want to know the generalisation error of the compressive network relative to the best achievable in the uncompressed class. The second difference is that rather than empirical average over the training points, here we have expectation w.r.t. the distribution of inputs. The final difference is the additional factor that represents the Lipschitz constant of the loss function.

Remarks

1. Let us relate this result to the analogous uncompressed neural network class $\mathcal{H}$. The first term on the r.h.s. is the best achievable generalisation error in

the original function class. The next two terms represent the complexity of the function class, which is reduced from $\sqrt{d}$ to $\sqrt{k}$ due to the dimensionality reduction. The price to pay for this reduction is a bias that we observe in the last two terms of the bound. This bias cannot be eliminated with more data as it is independent of N – instead it is governed by the compression dimension and scales as $\mathcal{O}(1/\sqrt{k})$. This scaling is the same as that of compressive OLS regression as known from previous work [8].

2. The quantities involved in the bound tell us about the characteristics of the original problem that influence the success of learning the network from compressed data. In particular, interesting to observe that this bound is not explicitly dependent on the network size M, so $M \to \infty$ is allowed provided that the norms involved in the bound are finite. The latter reflects regularity of the original problem. Similarly, the bound is independent of the original dimension of the data, d.
3. The bound holds under quite general conditions on R, however it is clear that a choice of Gaussian or subGaussian RP is preferable since then $\kappa_+ = 0$.
4. The choice of k balances between the distortion created by the RP (compressive distortion) and the complexity of the function class in the reduced space. We should also discuss the possibility of further reduction of the complexity term by means of constraining the norms of $(w_R)_i$. Indeed, reducing complexity by regularisation is well known from e.g. [3, 10]. However, such further constraint will introduce further bias in addition to the bias already made by the use of RP, which then would increase the magnitudes of the last two terms (while keeping their algebraic expression unchanged), and may be undesirable if k is small.
5. Finally, we should observe that the result does not require the input domain to be bounded. In case $\mathcal{X} \subseteq \mathbb{S}^{d-1}$, then of course the condition of finite expected norm is unnecessary.

In the light of this generalisation error, we may interpret our lower bound on the required k, eq. (16) as the degree of compressibility of the particular learning problem. That is, it tells us what makes a problem compressible. From it we can read off some precise conditions under which the function class can be learned from compressed data. Our bound suggests that it depends on the degree of regularity of the optimal learner in the original uncompressed function class, as expressed through the norms of the network's weights, and the Lipschitz constants.

In practice of course the quantities involved in the bound are unknown. There are other means to set k in practice, in particular existing model selection methods may be used (cross-validation, structural minimisation, etc). We do not pursue this here.

5 Conclusions

We gave a new lower bound on the required target dimension for compressed multi-layer perceptron to ensure small distortion of the outputs, which in contrast

with previous work, has no dependence on the network size or the sample size, and agrees with empirical behaviour. Using our findings, we briefly pursued a generalisation error analysis, which implies that compressed learning of the network has similar behaviour as the original in the sense that, for good generalisation, the size of the weights matters more than the size of the network.

Acknowledgement

The work of AK is funded by the EPSRC Fellowship EP/P004245/1.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-lindenstrauss with binary coins. j. Computer and System Sciences **66**(4), 671–687 (2003)
2. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. J. of the ACM **4**, 615–631 (1997)
3. Bartlett, P.L.: For valid generalization, the size of the weights is more important than the size of the network. In Neural Information Processing Systems **9**, 134–140 (1997)
4. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: Risk bounds and structural results. J. of Machine Learning Research **3**, 463–482 (2002)
5. Cybenko, G.: Approximations by superpositions of lipschitz continuous functions. Mathematics of Control, Signals, and Systems **2**(4), 303–314 (1989)
6. Dudley, R.M.: Uniform central limit theorems. Cambridge University Press, Cambridge, MA (1999)
7. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz maps into a Hilbert space. Contemporary Mathematics 26 pp. 189–206 (1984)
8. Kabán, A.: New bounds on compressed linear least squares regression. In: International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 448–456. JMLR W&P 33 (2014)
9. Kabán, A.: Improved bounds on the dot product under random projection and random sign projection. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). pp. 487–496 (2015)
10. Kakade, S.M., Sridharan, K., Tewari, A.: On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. in Neural Information Processing Systems (NIPS), pp pp. 793–800 (2008)
11. Kane, D.M., Nelson, J.: Sparser johnson-lindenstrauss transforms. J. of the ACM **61** (2014)
12. Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers, ann. Statist **30**(1), 1–50 (2002)
13. Reboredo, H., Renna, F., Calderbank, R., Rodrigues, M.R.D.: Bounds on the number of measurements for reliable compressive classification. IEEE Transactions on Signal Processing **64**(22), 5778–5793 (2016)
14. Skubalska-Rafajlowicz, E.: Neural networks with lipschitz continuous activation functions: dimension reduction using normal random projection. Nonlinear Analysis **71**, 12 (2009)